Semantic Segmentation Using Hybrid Markov Logic Networks

Aravindh Mahendran
amahend1@andrew.cmu.edu

Nitish Thatte
nitisht@andrew.cmu.edu

Adwait Gandhe
agandhe@andrew.cmu.edu

Abstract

Semantic image segmentation is the process of assigning human relevant labels to pixels in an image and is a high level vision problem. Increasing the number of labels increases the complexity of the problem. Markov Logic Networks (MLNs) allow us to handle uncertainty and complexity in a single framework, whereas Hybrid Markov Logic Networks (HMLNs) are an extension of the MLNs that allow continuous properties and functions over those properties as features. In this paper, we propose a method for semantic segmentation using Hybrid Markov Logic Networks, which integrate first order logic and statistical learning.

1 Introduction

Semantic segmentation is the process of assigning a class label to each pixel of the image. This is an important problem in computer vision for understanding the underlying information in an image. While classical segmentation techniques group together the pixels based on low level features, semantic segmentation adopts a supervised learning approach. There are two common approaches for semantic segmentation. The first makes use of low level features and combines them with a learning framework to obtain higher level labels. The second approach is to use low level cues, rather than features, with random fields and learn a unified framework using low level segmentation. In this paper we explore the second approach further by combining logical and statistical techniques to jointly address the issues of uncertainty and complexity.

This paper is structured as follows: Section 2 discusses the related work. In section 3, we discuss methods that we have attempted for semantic segmentation. The experiments conducted and the results obtained are presented in section 4. Section 5 presents our approach for the second half of the semester. We summarize our conclusions in section 6. And finally section A outlines our plan to implement the proposed approach.

2 Related Work

One of the first approaches for simultaneous object segmentation and recognition utilizes an Implicit Shape Model that integrates both capabilities into a common probabilistic framework [12]. Another approach used a generative model based on the bag of words representation for such simultaneous recognition and segmentation [4]. Furthermore, a method based on some of these approaches for scoring low-level patches according to their class relevance and propagating these posterior probabilities to pixels has been developed [6].

Several recent approaches use random fields to incorporate local cues and impart global control without implementing low level segmentation. For example, [10] proposes a Bayesian method for combining top-down and bottom-up cues. Conditional random fields have also been used for this purpose [11] [16]. Finally, Textonboost [18] is an approach to learning a discriminative model of object classes incorporating appearance, shape and context information efficiently. We suggest [1] for other semantic image segmentation approaches.

These approaches handle the complexity and uncertainty inherent in the structured inference problem in different ways. An alternate approach is to use Markov Logic Networks [7] [17] that attach weights to first-order formulae and view

them as templates for building Markov Networks. Hybrid Markov Logic Networks (HMLNs)[19] are an extension that allow for continuous properties to appear as features. In this paper we discuss our attempts to use HMLNs for semantic segmentation. A recent approach for semantic segmentation is to use a learned label transfer confidence function to propagate labels between neighboring superpixels based on the geodesic distance metric [5]. Labels are iteratively propagated across edges of a graph of connected superpixels based on geodesic distance. Another recent approach [2] learns a single random forest and incrementally adds context features derived from coarser levels. Unlike Textonboost [18], which learns two random forests, this approach models the dependencies between contextual and non-contextual features directly. Semantic image segmentation has also been approached using the bag of words model. In one of the recent approaches, [8], the authors develop a conditional random field model that combines dictionary learning, feature categorization (assigning key points to visual words) and image semantic segmentation into one framework.

3 Attempted Methods

3.1 Label Propagation

In section 2, we discussed a method of propagating labels on a graph of connected superpixels based on a learned label transfer confidence function. This method introduced us to two ideas: propagating labels rather than directly minimizing a cost function to arrive at a segmentation and working with graphs of related superpixels. Building on these ideas, we learn the weights for a hybrid Markov logic network in order to refine a prior expected segmentation obtained from a baseline algorithm. The method incorporates an undirected graph of connected superpixels that expresses where labels can propagate.

Using a MLN provides several advantages. First, MLNs allows us to perform metric learning by discovering a different weight for every label and feature vector component combination. Therefore, an MLN can model complex systems more easily than an SVM or logistic regression, which can learn one weight vector for all labels. Another advantage of MLNs is that they allow the user to express a template for his or her model using first order logic, a language that is easy to understand and learn.

Definition of Markov logic network: Numeric terms and predicates:

$\text{featureDistance}(s_i, s_j, f_m)$	Numeric term equal to the distance based on the m^{th} component of feature vectors for superpixels s_i and s_j.
$\text{isLabel}(s_i, l_k)$	Boolean predicate, True if s_i is assigned label l_k, false otherwise.
$\text{isNeighbor}(s_i, s_j)$	Boolean predicate, True if s_i and s_j are neighbors in our superpixel connectivity graph.

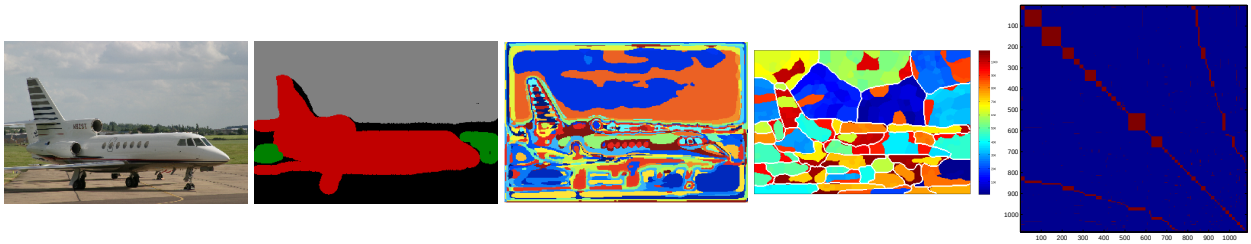
Formulae:

$$\text{isNeighbor}(s_i, s_j) \Rightarrow (\text{isLabel}(s_i, l_k) \Leftrightarrow \text{isLabel}(s_k, l_m))$$
$$(\text{isLabel}(s_i, l_k) \Leftrightarrow \text{isLabel}(s_k, l_m)) \times \text{featureDistance}(s_i, s_k, f_m)$$

These formulae encode that Superpixels with similar feature vectors should have the same label and superpixels that are connected by an edge (are neighbors) should also have the same label.

A brute-force approach to this problem would be to allow every superpixel to transfer its label to every other superpixel. Such a graph would allow superpixel labels to be learned from a more global context in the image. However, this would overly complicate the underlying Markov random field, thus making learning and inference intractable. In contrast, a simple approach is to connect nodes that are directly share borders. This approach, however, would only use local context for label propagation.

Therefore, we propose to take a middle ground approach that will produce densely connected superpixels where relations are expected to be strong, and fewer edges across boundaries expected not to share labels. In order to build this superpixel graph, we take two levels of superpixel segmentations: A coarse level, which we will henceforth refer to as supersuperpixels, and a fine level, which we term superpixels. These segmentations were found using the algorithm outlined in [13]. All superpixels within a supersuperpixel are connected allowing for contextual information to be shared within a region. Additionally, superpixels that share a boundary between two supersuperpixels are connected allowing for labels to propagate across supersuperpixels while acknowledging that label transfer across these borders is less probable than within a supersuperpixel. We then search through the training data set and select the $k-$Nearest Neighbors based on euclidean distance between GIST features [14] of the images. We use the superpixels in these k images to train the MLN on the formulae listed above.



(a) original image (b) ground truth segmentation (c) visual words map (d) superpixel segmentation (e) superpixel adjacency matrix

Figure 1: Pre-processing pipeline for label propagation experiment.

3.2 Other Approaches

We attempted to learn a HMLN on an ensemble of feature sets such as those used by [2]. In spite of the advantages of this approach, the HMLN does not afford the use of a large number of features the way random forests do. Hence we abandoned this approach. Even though the bag of words approach developed in [8] has shown very good results on the CAMVID[3] and Graz-02 [15] datasets, it cannot be incorporated into a MLN as it requires second order logic. Particularly, the word count histogram built as part of the bag of words model requires a count of the number of true predicates of a certain kind which is a meta level information beyond the first order logic. Hence, we abandoned this approach.

4 Experiment

We used the Alchemy - Open Source AI software [9] for learning the weights for the HMLN. Results described in this section are from the MSRC v1 dataset [20]. The 3 nearest neighbors from the training set of a query image are retrieved. Superpixels (fine and coarse, 1000 and 40 superpixels respectively), adjacency matrices and word maps are computed for each image. These are illustrated in 1. An initial baseline segmentation for the test image is computed by matching bag of words histograms of individual superpixels to histograms from the training set. The $28-$nearest histograms based on euclidean distance were retrieved and a probability vector was computed based on the counts of individual labels in this collection. Unfortunately, we ran into memory issues with the system using more than 23 Gb of virtual memory forcing it to kill the process.

5 New Proposed Method

After attempting such varied methods using Markov logic networks, we realize that Markov logic is not a good solution to problems in low level vision due to combinatorial explosion. Therefore, we propose to use Markov logic for high level reasoning only. We consider all possible outputs of a meanshift based unsupervised segmentation algorithm by varying the parameters involved and use Markov logic to choose the best of these predictions. We generate a random set of parameter values for the unsupervised segmentation algorithm. An SVM classifier is trained to predict the label for each segment. Each such semantic segmentation corresponds to a parse of the image. An HMLN associates a consistency value with a given parse which evaluates the initial segmentation and classifier output. We choose the best current parameter settings and repeat this process by generating another random set of parameters within the neighborhood of the current values.

Following the implementation of this approach, we shall experiment on the CAMVID dataset. This dataset consists of 4 video sequences. In the first set of experiments, we shall train on a random subset of frames in each video and test the trained model on the remaining frames of that video itself. The split would be 50:50. Parameters, specifically the window size and sample resolution for meanshift parameter generation, will be derived using cross validation. A second set of experiments would train the model on one video sequence and test it on the other three.

6 Conclusion

As mentioned in the project proposal, we have attempted to model different approaches for semantic image segmentation using HMLNs. We further analyzed why these are infeasible and proposed a new method for addressing the problem.

References

[1] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *In CVPR*, 2012.

[2] Erik Rodner Bjrn Frhlich and Joachim Denzler. Semantic segmentation with millions of features: Integrating multiple cues in a combined random forest approach. In *In ACCV*, 2012.

[3] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, xx(x):xx–xx, 2008.

[4] Liangliang Cao and Fei-Fei Li. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *ICCV'07*, pages 1–8, 2007.

[5] Xiaowu Chen, Qing Li, Yafei Song, Xin Jin, and Qinping Zhao. Supervised geodesic propagation for semantic label transfer. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision ECCV 2012*, volume 7574 of *Lecture Notes in Computer Science*, pages 553–565. Springer Berlin / Heidelberg, 2012.

[6] Gabriela Csurka and Florent Perronnin. A simple high performance approach to semantic segmentation. In Mark Everingham, Chris J. Needham, and Roberto Fraile, editors, *BMVC*. British Machine Vision Association, 2008.

[7] Pedro Domingos, Stanley Kok, Hoifung Poon, Matthew Richardson, and Parag Singla. Unifying logical and statistical ai. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 2–7. AAAI Press, 2006.

[8] Aastha Jain, Luca Zappella, Patrick McClure, and Ren Vidal. Visual dictionary learning for joint object categorization and segmentation. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision ECCV 2012*, volume 7576 of *Lecture Notes in Computer Science*, pages 718–731. Springer Berlin Heidelberg, 2012.

[9] Stanley Kok, Marc Sumner, Matthew Richardson, Parag Singla, Hoifung Poon, Daniel Lowd, Jue Wang, Aniruddh Nath, and Pedro Domingos. The alchemy system for statistical relational ai. Technical report, Department of Computer Science and Engineering, University of Washington, Seattle, WA, 2010.

[10] M. Pawan Kumar, Philip H. S. Torr, and A. Zisserman. Obj cut. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 18–25, Washington, DC, USA, 2005. IEEE Computer Society.

[11] Sanjiv Kumar and Martial Hebert. A hierarchical field framework for unified context-based classification. In *Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2*, ICCV '05, pages 1284–1291, Washington, DC, USA, 2005. IEEE Computer Society.

[12] Bastian Leibe, Ales Leonardis, and Bernt Schiele. Combined object categorization and segmentation with an implicit shape model. In *In ECCV workshop on statistical learning in computer vision*, pages 17–32, 2004.

[13] G. Mori. Guiding model search using segmentation. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1417 –1423 Vol. 2, oct. 2005.

[14] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

[15] Opelt, A., and A Pinz. The tu graz-02 database. Technical report, TU Graz, 2002.

[16] Xuming He Richard, Richard S. Zemel, and Miguel . Carreira-perpin. Multiscale conditional random fields for image labeling. In *In CVPR*, pages 695–702, 2004.

[17] Matthew Richardson and Pedro Domingos. Markov logic networks. In *Machine Learning*, page 2006, 2006.

[18] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *In ECCV*, pages 1–15, 2006.

[19] Jue Wang and Pedro Domingos. Hybrid markov logic networks.

[20] John M. Winn, Antonio Criminisi, and Thomas P. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, pages 1800–1807, 2005.

A Plan of Activities

Old Plan of Activities

From now until the midterm report, Nitish will work on writing HMLNs, Aravindh will work on feature extraction and Adwait will work on super-pixel generation techniques. The midterm report will consist of the results of initial experiments with different techniques for each aspect and present our design choices.

New Plan of Activities

Work so far suggest a radical change in approach. Aravindh will work on meanshift segmentation and understand its parameter space. Nitish and Adwait will work on building HMLNs for tree parsing and consistency evaluation.